**PART A**

**Multiple Choice Questions**

**Time: 90 minutes**

**1). In K-fold cross validation, finally which model is chosen over?**

1. Model with smallest validation accuracy
2. Model with highest average validation accuracy
3. Model with smallest train accuracy
4. Model with smallest train accuracy

**2). When you perform K fold validation on a dataset which has 1000 observations with k = 100. Each fold will have how many data samples?**

a)100

b)1

c)1000

d)10

**3). What is the minimum value of ‘k’ to perform K-fold cross validation?**

a)1

b)2

c)4

d)3

**4). In a city of 1 million, 500 people have been diagnosed with cancer, whereas the rest of the people do not have cancer according to a recent medical study. Such a class distribution is considered to be**

1. Balanced
2. Imbalanced

**5). ‘Leave one out cross validation' is a special case of K-fold cross validation with K =**

a) no. of observations/2

b)1

c) 0

d) no.of. observations

**6). If a model \_\_\_\_\_\_\_\_\_ the training data, it is considered to have high bias and low variance.**

a) overfits

b) underfits

c)perfectly fits

**7). To mimic the performance of a model in production, we predict labels and assess the model using which of the following**

a) train data

b) test data

c)validation data

**8). Which of the following techniques can be employed to balance a dataset with imbalanced classes?**

a) Cross Validation

b) SMOTE

c) Upsampling and downsampling

d) Both b and c

**9). SMOTE (Synthetic Minority Over-sampling Technique) uses which of the following algorithms to create synthetic data**

1. KNN
2. Decision Trees
3. Logistic Regression
4. Linear Regression

**10). When you perform a linear regression and the model is said to have high variance and low bias. How do you improve the performance of the current model?**

a) Use regularization methods to reduce overfitting

b) Use a sophisticated model like SVM

c)Perform cross validation with different values of k

d) Change the split of training and test data

**11). What do we have on the X-axis of an ROC graph?**

a) FPR

b) TNR

c) FNR

d) TPR

**12). Let’s say we are using Grid search CV to find out the optimal values of k in KNN algorithm to find the best accuracy score. Which of the below algorithms Grid search uses to find out the accuracy score in the above case?**

a) Stratified sampling

b) Bootstrap sampling

c) Leave one out cross validation

d) Cross validation

**13). Which of the following is not true in the case of Randomized Search CV?**

a). It converges comparatively faster than GridSearch CV

b) t doesn’t guarantee to give the best parameters combination.

c) It uses random combinations of the hyperparameters to find the best solution for the built model.

d) It always gives the best parameters combination for a given model.

**14) Which of the following technique creates synthetic data points?**

a) SMOTE

b) KNN

c) Stratified down-sampling

d) Stratified up-sampling

**15) Which of the given algorithm tries every possible combination of a list of values of the hyper-parameters and choose the best combination based on the cross-validation score?**

a) Randomized Search CV

b) GridSearch CV

**16) Which of the following is not considered as Hyperparameter in a Machine Learning Model?**

a) Shrinkage factor in Ridge

b) Number of clusters in K Means

c)Depth of a tree in a Decision Tree

d) Coefficients in a linear regression model

**17) Which of the parameters are considered to be hyper parameters for Support vector machines?**

a) Gamma

b) Distance between the hyperplane and support vectors

c) Support vectors

d) None

**18) Which of these is/are Hyperparameter Optimization method?**

a) Grid Search

b) Random search

c) K-fold cross validation

d)Both a and b

**19) Which of the following is not an example of grid search?**

a) Training a SVM model with different combinations of ‘c’ and ‘gamma’

b) To select best model for the given statement

c) To evaluate Ridge regression algorithm, we train the model with different alpha values

d)To get the best hyperparameters (like max\_depth) for Random Forest model?

**20). Suppose you have picked the parameter for a model using 10-fold cross validation (CV). Which of the following is the best way to pick a final model to use and estimate its error?**

a) Pick any of the 10 models you built for your model; use its error estimate on the held-out data

b) Train a new model on the full data set, using the parameter you found; use the average CV error as its error estimate

c) Average all of the 10 models you got; use the average CV error as its error estimate

d) Average all of the 10 models you got; use the error the combined model gives on the full training set

**21). Which of the following statements is true about the learning rate alpha in gradient descent?**

a) If alpha is very small, gradient descent will be fast to converge. If alpha is too large, gradient descent will overshoot

b) If alpha is very small, gradient descent can be slow to converge. If alpha is too large, gradient descent will overshoot

c) If alpha is very small, gradient descent can be slow to converge. If alpha is too large, gradient descent can be slow too

d) If alpha is very small, gradient descent will be fast to converge. If alpha is too large, gradient descent will be slow

## 22). **How to select best hyperparameters in tree-based models?**

1. measure performance over training data
2. measure performance over validation data
3. both of these
4. random selection of hyper parameters

**23). In this Linear Regression Equation: y = -3.4x - 2.5, What is the coefficient, dependent variable, independent variable and constant respectively?**

a) x, y, -2.5, -3.4

b) x, -3.4, -2.5, y

c) -3.4, y, x, -2.5

d) -3.4, -2.5, x, y

**24). When you predict the output of a variable based on a set of labelled data, this is what form of machine learning?**

a) Reinforcement Learning

b) Supervised Learning

c) Unsupervised Learning

**25) Which of the below is a practical application of Supervised Learning?**

a) Predicting the amount of sales based on the past customer data

b) Reducing number of Dimensions

c) Dividing customer into 4 different segments.

d) Making a machine to learn to play a competitive chess game

**26) If, 'w','x','y','z' are the predictors and 't' is the target, from the equation,**

**t = 10 + 0.3w + 0.5x + 0.2y + z**

**If all the predictors are zero except 'z' and z = 8, then t** =

1. 18
2. 36
3. 9
4. 27

**27). The objective of Linear regression is to (d1 + d2 + d3 + d4) where d1, d2, d3 and d4 represents the sum of squared vertical distances between observed and predicted values respectively.**

a) Maximise the error

b) Minimise the error

**28). What is the objective of the Simple Linear Regression algorithm?**

a) To maximize the average sum of squared errors

b) To obtain a line that best fits the data

c) To find linear relationship between independent variables.

d) All the above

**29). A simple linear regression line for the set of n data points is given by the equation of a line. The equation of the same line in slope intercept form is**

a) ax+by+cd

b) ax+b

c) y= ax2 + bx+ c

d) None of these

**30). Logistic Regression is a classification algorithm which:**

a) Calculates probabilities using a sigmoid function for the target variable and converts into labels based on a threshold.

b) Calculates probabilities for the target variable using the equation of a line and converts into labels based on a threshold.

c) Directly calculates the labels of the target variable without calculating the probability

**31). Which of the following evaluation metric can be used to access a Logistic Regression model?**

a) MAPE

b) Precision

c) RMSE

d) R square

**32). Logit function takes an input variable which can range from (-LaTeX: \infty to +LaTeX: \infty) and maps it to [x,y] where x and y are**

a) -1 and 1

b) 1 and infinity

c) 0 and 1

d)-1 and 0

**33). The confusion matrix is used to**

a) How data is spread on each dimension

b) Evaluate the performance of classification algorithms within each class

c) Find the correlation between different attributes

d) Understand how attributes are related to each other

**34) Which of the following is not true about precision and recall?**

a) Recall = 'True Positive' / (True Positive + False Negative)

b) Precision = 'True Positive' / (False Positive + False Negative)

c) Precision and recall are class level metrics in classification problems.

d) Precision = 'True Positive' / (True Positive + False Positive)

**35). What is the loss function used in Logistic Regression to find the best fit line?**

a) MSE

b) MAE

c) Log-Loss

d)MAPE

**36). Increasing the threshold does not change the values in the confusion matrix of a model for a given dataset.**

a) True

b) False

**37). Which of these is a valid kernel input for the kernel parameter in sklearn.svm.SVC()?**

a) rbf

b) sigmoid

c) poly

d) all of the above

**38). Kernel trick is used because-**

a) It offers complex way to scale data to higher dimension

b) It offers computationally inexpensive way to scale data into higher dimensions

c) It offers data augmentation to better fit our data

d) None of these

**39). Increasing gamma and c simultaneously, generally initiates this transition in terms of the model:**

a) Underfit to Overfit

b) Overfit to underfit

**40). Linear SVMs have no hyper parameters that need to be set by cross-validation.**

a) True

b) False

**41). Increasing 'c', \_\_\_\_\_\_\_\_\_ the  margin width.**

a) increases

b) shrinks

c)Merges

d)Dissolves

**42). What is entropy in Decision Tree?**

a) Measure of predictability

b) Measure of Bias

c) Measure of disorder

d) None

**43). Which of the following nodes have the maximum entropy in a decision tree?**

a) Leaf Node

b) Terminal node

c) Root node

d) Decision node

**44). Which of the following parameters (in python) controls the number of samples in the leaf node?**

a) max\_features

b) min\_samples

c) random\_state

d) max\_depth

**45). Which ensemble method creates several subsets of data from original data with replacement and aggregate the final results by combining the results from training different subsets?**

a) Bagging

b) Stacking

c) Boosting

d) Random Forest

**(46). Ensemble techniques leverage the fact that collective wisdom is usually better than individual wisdom**

a) True

b) False

**47). What will be the Euclidean distance of A(4,0) be from the centroid of cluster which has 2 data points.  
C1: {(3,3),(5,5)}**

a) 3

b)2

c) 16

d) 4

**48). Which of the following is not true in the case of k means clustering?**

a) The data points that are the farthest to a centroid will create a cluster

b) Choosing different starting points can result in different clusters

c) It requires the number of clusters to be specified

d) k means clusters data by separating data points into groups of equal variance

**49). Covariance matrix is a mathematical representation of**

a) Variance of individual dimensions and pairs of dimensions

b)  Sum of Total Variance in the individual dimensions and across dimensions

c) Coefficient of variance of individual dimensions

**50) What does measuring distance between clusters mean in case of complete linkage?**

a) Minimum Distance between pair of records in cluster A and B respectively

b) Maximum Distance between pair of records in cluster A and B respectively

c) Distance between centroids of 2 different clusters A and B

d) Average distance of all possible distances between records in one cluster to records in other cluster

**PART B**

Given is the ‘Portugal Bank Marketing’ dataset: Bank client data:

1) age (numeric)

2) job: type of job(categorical:"admin.","bluecollar","entrepreneur","housemaid","management","retired","selfemployed","services","student","technician","unemployed","unknown")

3) marital: marital status (categorical: "divorced","married","single","unknown"; note: "divorced" means divorced or widowed)

4) education: education of individual (categorical: "basic.4y","basic.6y","basic.9y","high. school","illiterate","professional. course","university. degree","unknown")

5) default: has credit in default? (Categorical: "no","yes","unknown")

6) housing: has housing loan? (categorical: "no","yes","unknown")

7) loan: has personal loan? (categorical: "no","yes","unknown") Related with the last contact of the current campaign:

8) contact: contact communication type (categorical:"cellular","telephone")

9) month: last contact month of year (categorical: "jan", "feb", "mar", …, "nov", "dec")

10) dayofweek: last contact day of the week (categorical: "mon","tue","wed","thu","fri")

11) duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y="no"). Yet, the duration is not known before a call is performed. Also, after the end of the call y is obviously known. Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to have a realistic predictive model. Exam Paper 1 Other attributes:

12) campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)

13) pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)

14) previous: number of contacts performed before this campaign and for this client (numeric)

15) poutcome: outcome of the previous marketing campaign (categorical: "failure","nonexistent","success") Social and economic context attributes

16) emp.var.rate: employment variation rate - quarterly indicator (numeric)

17) cons.price.idx: consumer price index - monthly indicator (numeric)

18) cons.conf.idx: consumer confidence index - monthly indicator (numeric)

19) concavepoints\_se: standard error for number of concave portions of the contour

20) euribor3m: euribor 3-month rate - daily indicator (numeric)

21) nr.employed: number of employees - quarterly indicator (numeric) Output variable (desired target):

22) y: has the client subscribed a term deposit? (binary: "yes","no") Perform the following tasks: Marks

Q1. What does the primary analysis of several categorical features reveal? [5]

Q2. Perform the following Exploratory Data Analysis tasks: a. Missing Value Analysis b. Label Encoding wherever required c. Selecting important features based on Random Forest d. Handling unbalanced data using SMOTE e. Standardize the data using the anyone of the scalers provided by sklearn [10] Exam Paper 1

Q3. Build the following Supervised Learning models: a. Logistic Regression b. AdaBoost c. Naïve Bayes d. KNN e. SVM XGBoost g.decision tree h.deeplearning[10] Note that none of the models should overfit

Q4. Tabulate the performance metrics of all the above models and tell which model performs better in predicting if the client will subscribe to term deposit or not